Contrastive Learning can Identify the Underlying Generative Factors of the Data

Roland S. Zimmermann[1,2*] **Steffen Schneider**[1,2*] **Yash Sharma**[1,2*]
Matthias Bethge[1†] **Wieland Brendel**[1†]

[1]University of Tübingen [2]IMPRS for Intelligent Systems

Abstract

Contrastive learning has recently seen tremendous success in unsupervised learning, but the understanding of the source of their effective generalization to a large variety of downstream tasks has been limited. We rigorously show that feedforward models trained on a common contrastive loss can implicitly invert the underlying generative model of the observed data up to affine transformations. While we detail the set of assumptions which need to be met to prove this result, our empirical results suggest our findings are robust to considerable model mismatch. We demonstrate contrastive learning performs comparably to the state-of-the-art in disentanglement on benchmark datasets, a notable observation due to the unique lack of an explicit generative objective. This highlights a deep connection between contrastive learning, generative modeling, and nonlinear independent component analysis, providing a theoretical foundation to derive more effective contrastive losses while simultaneously furthering our understanding of the learned representations.

1 Introduction and Related Work

Contrastive learning has been tremendously successful in unsupervised representation learning for image and sequential data [1–13]. In essence, contrastive methods aim to learn representations where related samples are aligned (positive pairs, e.g. augmentations of the same image), while unrelated samples are separated (negative pairs) [9]. This intuitively leads to invariance to irrelevant details or transformations (by decreasing the distance between positive pairs), while preserving a sufficient amount of information about the input needed for solving downstream tasks (by increasing the distance between negative pairs) [14].

In this work, we move beyond intuition by showing that the encoder learned with a contrastive loss can recover the true generative factors of variation (up to affine transformations) if the true generative model of the data matches the assumptions made by the model. This theory bridges the gap between contrastive learning, and the fields of nonlinear ICA and generative modeling. We verify our theoretical findings with controlled experiments and provide evidence that our theory holds true in practice, even if the assumptions on the ground-truth generative model are partially violated.

Contrastive Learning Despite the success of contrastive learning (CL), the theoretical understanding remains limited. One way to theoretically motivate CL is to refer to the InfoMax principle [15], which corresponds to maximizing the mutual information (MI) between different views [3, 6, 7, 9, 14]. However, as optimizing a tighter bound on the MI can produce worse representations [16], it is not clear how accurate this motivation describes the behavior of CL. Another approach aims to explain the success by the introducing latent classes [17]. While this theory has some appeal, there exists

*Equal contribution. †Joint supervision. Correspondence: first.last@bethgelab.org

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

a gap between empirical observations and its predictions, namely that using an excessive number of negative samples decreases the performance [2, 5, 8, 9]. More recently, the behavior of CL has been analyzed from the perspective of *alignment* and *uniformity* properties of representations, by demonstrating a correlation between both metrics and downstream task performance on representation learning tasks[18]. We build on these results to make a connection to cross-entropy minimization, leveraging this to provide identifiability results for a practically successful instantiation of CL.

Nonlinear ICA Nonlinear Independent Components Analysis (ICA) attempts to find the underlying components for multidimensional data. Said components correspond to a well-defined generative model g, which is assumed to be invertible [19, 20]. In other words, nonlinear ICA solves a demixing problem, i.e., given observed data $\mathbf{x} = g(\mathbf{s})$, it finds the inverse model g^{-1} that allows the original sources $\mathbf{s}$ to be recovered. Hyvärinen et al. [21] show that the nonlinear demixing problem can be solved as long as the independent components are conditionally mutually independent with respect to some auxiliary variable. The authors further provide practical estimation algorithms for solving the nonlinear ICA problem [22–24]. Khemakhem et al. [25] show that this framework generalizes to a broad family of deep latent-variable models. While contrastive learning has previously been used for estimation of specified probabilistic models [21, 23, 24], our work instead leverages identifiability as a tool for understanding instantiations found to be practically successful for representing high-dimensional, complex sensory input [9]. In a similar vein, [26] provide identifiability conditions satisfied by practically successful models for representation learning, showing different representation functions, learned on the same data distribution, live within linear transformations of each other. We instead analyze the relation with respect to the data generating process, showing the inverse of that process lives within a linear transformation of the contrastive optimum.

2 Theory

We will show a connection between contrastive learning and identifiability in the form of nonlinear ICA. For this, we introduce a feature encoder f that maps observations $\mathbf{x}$ to representations. We consider the popular *InfoNCE* loss, which assumes ℓ_2 normalized representations, to perform CL

$$\mathcal{L}_{\mathsf{Contr}}(f;\tau,M) \quad := \quad \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathrm{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathrm{data}}}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau}}{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau} + \sum_{i=1}^M e^{f(\mathbf{x}_i^-)^\top f(\tilde{\mathbf{x}})/\tau}} \right]. \tag{1}$$

Here $M \in \mathbb{Z}_+$ is a fixed number of negative samples, p_{data} is the distribution of all observations and p_{pos} is the distribution of positive pairs. This loss is based on the InfoMax principle [15] and has been shown to be effective by many recent representation learning methods [1, 2, 5–9, 12]. Note that the theoretical results of this paper also hold for a loss function whose denominator only consists of the second summand (i.e. the SimCLR loss [9]).

In the spirit of literature on nonlinear ICA [21–25, 27–29], we assume that the observations $\mathbf{x} \in \mathcal{X}$ are generated by an invertible (injective) generative process $g : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^N$ is the space of observations and $\mathcal{Z}$ denotes the space of latent factors. Influenced by the feature normalization in InfoNCE, we further assume that $\mathcal{Z}$ is the unit hypersphere $\mathbb{S}^{N-1}$ (see Appx. A.1). Additionally, we assume that the ground-truth marginal distribution of the latents of the generative process is uniform and that the conditional distribution is a von Mises-Fisher (vMF) distribution:

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \qquad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \quad C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}} = \text{const.}, \quad \mathbf{x} = g(\mathbf{z}) \tag{2}$$

Given these assumptions, we will show that if f optimizes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$, then f solves the demixing problem, i.e., inverts g. Detailed proofs are given in Appx. A.3. Note that one can derive similar properties beyond spherical spaces, e.g. for $\mathbb{R}^N$ or convex bodies (Appx. A.4).

2.1 Relation between contrastive learning and cross-entropy minimization

From the perspective of nonlinear ICA, we are interested in understanding how the representations $f(\mathbf{x})$ which minimize the contrastive loss (1) are related to the ground-truth source signals $\mathbf{z}$. To study this relationship, we focus on the mapping $h = f \circ g$ between the recovered source signals $h(\mathbf{z})$ and the true source signals $\mathbf{z}$. A core insight is a connection between the contrastive loss and the

cross-entropy between the ground-truth latent distribution and a certain model distribution. For this, we expand the theoretical results obtained by Wang and Isola [18]:

Theorem 1 ($\mathcal{L}_{\text{contr}}$ converges to the cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\text{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{3}$$

where H is the cross-entropy between the ground-truth conditional distribution over positive pairs p and the conditional distribution over the recovered latent space q_{h}, and $C_h(\tilde{\mathbf{z}}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appx. A.2):

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\tilde{\mathbf{z}})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \,\mathrm{d}\mathbf{z}. \tag{4}$$

This result makes analyzing the problem of CL simpler, as it reduces it to deducing properties from the well-understood cross-entropy objective. Interestingly, for uniform ground truth marginal distributions on bounded spaces, $C_h^{-1}(\mathbf{z})$ also coincides with the pushforward of $p(\mathbf{z})$ through h (see Sec. A.5), which eventually becomes a uniform distribution at the optimum.

Next, we show that the minimizers h^* of the cross-entropy (4) are isometries in the sense that $\kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}})$ for all $\mathbf{z}$ and $\tilde{\mathbf{z}}$. In other words, they preserve the dot product between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} \exp(\kappa\tilde{\mathbf{z}}^\top\mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau/\kappa}$.*[2] *Consider the model conditional distribution*

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \quad \textit{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \,\mathrm{d}\tilde{\mathbf{z}}, \tag{5}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.

We show in the Appendix (Lemma 3) that this implies that the empirical marginal distribution (pushforward of p through h, denoted as $p_{\#h}$, which coincides with $C_h(\mathbf{z})^{-1}$) is also a uniform distribution, i.e., $p_{\#h}(\mathbf{z}) = \text{const}$.

2.2 Contrastive learning identifies the ground-truth factors

From the strong property of isometry, we can now deduce a key property of the minimizers h^*:

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$. If $h : \mathcal{Z} \to \mathcal{Z}$ maintains the dot product up to a constant factor, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$: $\kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an affine transformation.*

In the last step, we can combine the previous propositions to derive our main result: the minimizers of $\mathcal{L}_{\text{contr}}$ solve the demixing problem of nonlinear ICA up to linear transformations, i.e., they identify the original sources $\mathbf{z}$ for observations $g(\mathbf{z})$ up to linear transformations.

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). If the mixing function g is differentiable and invertible, and if f is differentiable and minimizes the CL loss (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is affine, i.e., it recovers the latent sources up to affine transformations.*

Note that we do not assume knowledge of the ground-truth generative model g; we only make assumptions about the distribution of latents and that f minimizes the contrastive loss. In Appx. Theorem 6, we show a similar result for $\mathcal{Z} = \mathbb{R}^N$ and a wider class of ground-truth conditional densities: if f minimizes the cross-entropy, then h is again an affine transformation. Having an exact match between the assumed model distribution q_{h} and the ground-truth conditional is unlikely to happen in practice. While a theoretical account for these cases is beyond the scope of this work, we provide empirical evidence that h is still an affine transformation even if there is a severe mismatch.

[2] Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Table 1: Robustness of CL to a mismatch between model assumptions and the ground truth, averaged over 5 runs. Note that only the first two rows correspond to settings that match our assumptions, while the others show results for violated assumptions (see column *M.*).

Generative process g			Model f			R^2 [%]		
Space	$p(\cdot)$	$p(\cdot\|\cdot)$	Space	$q_{\mathrm{h}}(\cdot\|\cdot)$	M.	Linear Score	Supervised Score	Unsupervised Score
$\mathbb{S}_1^9$	Uniform	vMF(κ=1)	$\mathbb{S}_1^9$	vMF(κ=1)	✓	77.58 ± 3.06	99.82 ± 0.03	99.42 ± 0.05
$[0,1]^{10}$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=2)	✓	87.71 ± 4.64	99.80 ± 0.03	99.52 ± 0.02
$\mathbb{S}_1^9$	Uniform	vMF(κ=10)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.58 ± 3.06	99.81 ± 0.03	99.86 ± 0.01
$\mathbb{S}_1^9$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.98 ± 2.97	99.81 ± 0.03	99.88 ± 0.03
$\mathbb{S}_1^9$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.58 ± 3.06	99.80 ± 0.04	99.86 ± 0.00
$[0,1]^{10}$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	87.71 ± 4.64	99.81 ± 0.03	99.53 ± 0.02
$[0,1]^{10}$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=3)	✗	87.71 ± 4.64	99.81 ± 0.03	99.70 ± 0.02
$[0,1]^{10}$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=3)	✗	87.71 ± 4.64	99.83 ± 0.03	99.69 ± 0.02
$\mathbb{S}_1^9$	GenNorm(β=2, λ=1)	GenNorm(β=1, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	78.34 ± 3.83	99.78 ± 0.04	98.94 ± 0.03
$\mathbb{S}_1^9$	GenNorm(β=2, λ=1)	GenNorm(β=2, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	78.05 ± 3.17	99.78 ± 0.05	98.94 ± 0.02
$\mathbb{R}^{10}$	GenNorm(β=1, λ=1)	GenNorm(β=2, λ=1)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	73.28 ± 2.35	99.76 ± 0.02	97.57 ± 0.37
$\mathbb{R}^{10}$	GenNorm(β=2, λ=1)	GenNorm(β=2, λ=1)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	75.43 ± 2.68	99.80 ± 0.04	98.70 ± 0.04

3 Experiments

Validation of theoretical claim We now validate our theoretical claims under both perfectly matching and violated conditions regarding the ground truth marginal distribution, the mixing function, and the ground truth conditional. We consider source signals of dimensionality $n = 10$. We sample pairs of source signals in two steps: First, we sample from $p(\mathbf{z})$, which, if chosen to be uniform, matches our assumptions, if not (e.g. Normal distribution), violates our assumptions. Second, we generate the positive pair by sampling from a conditional $p(\tilde{\mathbf{z}}|\mathbf{z})$. We also consider spaces beyond the hypersphere, such as unbounded $\mathbb{R}^N$ and the bounded hypercube (which is a convex body). We generate the observations with a multi-layer perceptron (MLP) following the settings used by [23, 24]. Specifically, we use leaky ReLU units and control the condition number of the weight matrix to ensure that the MLP is invertible. For more details about the experimental setup and used loss functions, see Sec. A.6.

To test for identifiability up to affine transformations, we fit a linear regression between the ground-truth and recovered sources and report the coefficient of determination (R^2). In each setting, we report three scores. First, we ensure that the problem requires nonlinear demixing by considering a linear model, which amounts to the score of a linear fit between observations and sources (Linear Score). Second, we ensure that the problem is solvable within our model class by training our model f with supervision, minimizing the mean-squared error between $f(g(\mathbf{z}))$ and $\mathbf{z}$ (Supervised Score; upper bound). Third, we fit our model without supervision using a contrastive loss (Unsupervised Score). The results in Table 1 show that CL recovers a score close to the empirical upper bound, and mismatch in assumptions on the marginal and conditional only lead to a slight drop in performance.

Extensions to image data We evaluate CL on disentanglement with complex pixel inputs (Appx. A.6.2, 30, 31), benchmarking applicability beyond the theoretical conditions. For better control over the objective, we split the CL loss into uniformity and alignment terms (Appx. A.6.1). Appx. Table 2 shows that scores increase consistently with an increased relative weight on the uniformity term, denoting the importance of uniformity in preventing collapse due to optimization of the alignment loss. For Laplace transitions of the latents (LAP), the BetaVAE score [32], which corresponds to fitting a logistic classifier to the absolute differences in the model latents, reaches 100%, providing evidence to the robustness of CL in identifying the sources up to affine transformations. Finally, we note the surprising performance on KITTI Masks (Appx. Table 5), which contains natural shapes and continuous natural transitions, hinting at the scalability benefits of CL.

4 Conclusion

This work shows that contrastive learning can uncover the true generative factors of variation underlying the observational data. We verify this claim theoretically and give evidence that our theory also holds under much milder assumptions on the generative model. Our work connects CL, generative modeling and nonlinear ICA, thereby laying a strong theoretical foundation for one of the most successful self-supervised learning techniques.

References

[1] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

[2] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[4] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[5] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[6] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[7] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15509–15519, 2019.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[10] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

[11] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *CoRR*, abs/1910.05453, 2020.

[12] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, 2020.

[13] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE, 2020.

[14] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning, 2020.

[15] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[16] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[17] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019.

[18] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

[19] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[20] Christian Jutten, Massoud Babaie-Zadeh, and Juha Karhunen. Nonlinear mixtures. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pages 549–592, 2010.

[21] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

[22] Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

[23] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

[24] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Proceedings of Machine Learning Research*, 2017.

[25] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[26] Geoffrey Roeder, Luke Metz, and Diedrik P. Kingma. On linear identifiability of learned representations. *arXiv preprint arXiv:2007.00810*, 2020.

[27] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[28] Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, and Klaus-Robert Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

[29] Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

[30] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

[31] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:xxx*, 2020.

[32] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2(5):6, 2017.

[33] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances, 2017.

[34] I. Ahmad and Pi-Erh Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

[35] Michael Ruzhansky and Mitsuru Sugimoto. On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences*, 5(1):13–18, 2015.

[36] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 606–607. Springer, 2013.

[37] Stanisław Mazur and Stanisław Ulam. Sur les transformations isométriques d'espaces vectoriels normés. *CR Acad. Sci. Paris*, 194(946-948):116, 1932.

[38] Bogdan Nica. The mazur-ulam theorem, 2013.

[39] Piotr Mankiewicz. Extension of isometries in normed linear spaces. *Bulletin de l'Academie polonaise des sciences: Serie des sciences mathematiques, astronomiques et physiques*, 20(5): 367–+, 1972.

[40] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[41] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[42] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *arXiv preprint arXiv:1905.04804*, 2019.

[43] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[44] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

A Appendix

A.1 Assumptions: generative process

Let the generator $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{X} \subseteq \mathbb{R}^N$. We assume that the marginal distribution $p(\mathbf{z})$ over latent variables $\mathbf{z} \in \mathcal{Z}$ is uniform:

$$p(\mathbf{z}) = \frac{1}{|\mathcal{Z}|}. \tag{6}$$

Further, we assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is a von Mises-Fisher distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \quad C_p := \int e^{\kappa \boldsymbol{\eta}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{7}$$

where κ is a parameter controlled the width of the distribution and $\boldsymbol{\eta}$ is any vector on the hypersphere. Finally, we assume that during training one has access to samples from both of these distributions.

A.2 Assumptions: model

Let $f : \mathcal{X} \to \mathbb{S}_r^{N-1}$, where $\mathbb{S}_r^{N-1}$ denotes a hypersphere with radius r, be the model whose parameters are optimized using constrastive learning. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through $h = f \circ g$ and

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \text{with} \quad C_q(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top \, h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{8}$$

where $C_q(\mathbf{z})$ is the partition function and $\tau > 0$.

A.3 Proofs for Sec. 3

We begin be recalling a result of Wang and Isola [18], where the authors show an asymptotic relation between the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ and two loss functions, the *alignment* loss $\mathcal{L}_{\mathsf{align}}$ and the *uniformity* loss $\mathcal{L}_{\mathsf{uniform}}$:

Proposition A (Asymptotics of $\mathcal{L}_{\mathsf{contr}}$, 18)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \to \infty} \mathcal{L}_{\mathsf{contr}}(f; \tau, M) - \log M = \mathcal{L}_{\mathsf{align}}(f; \tau) + \mathcal{L}_{\mathsf{uniform}}(f; \tau), \tag{9}$$

where

$$\begin{aligned} \mathcal{L}_{\mathsf{align}}(f; \tau) &:= -\frac{1}{\tau} \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}}, \mathbf{z}) \sim p(\tilde{\mathbf{z}}, \mathbf{z})} \left[(f \circ g)(\mathbf{z})^\mathsf{T} (f \circ g)(\mathbf{z}) \right] \\ \mathcal{L}_{\mathsf{uniform}}(f; \tau) &:= \mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{(f \circ g)(\tilde{\mathbf{z}})^\mathsf{T} (f \circ g)(\mathbf{z})/\tau} \right] \right]. \end{aligned} \tag{10}$$

Proof. See Wang and Isola [18]. Note that they originally formulated the losses in terms of observations $\mathbf{x}$ and not in terms of the latent variables $\mathbf{z}$. However, this modified version simplifies notation in the following. □

Based on this result, we show that the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ asymptotically converges to the cross-entropy between the ground-truth conditional p and our assumed model conditional distribution q_h. This is notable, because given the correct model specification for q_h, the cross entropy is minimized iff $q_h = p$, i.e., the ground truth conditional distribution and the model distribution will match.

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ converges to cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \to \infty} \mathcal{L}_{\mathsf{contr}}(f; \tau, M) - \log M + \log |\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z})) \right] \tag{11}$$

where H is the cross-entropy between the ground-truth conditional distribution over positive pairs p and the conditional distribution over the recovered latent space q_{h}, and $C_h(\tilde{\mathbf{z}}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appendix A.2):

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\tilde{\mathbf{z}})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\mathbf{z}. \tag{12}$$

Proof. The cross-entropy between the conditional distributions p and q_h is given by

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z})) \right] \tag{13}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z}) \right] \right] \tag{14}$$

$$= \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[-\frac{1}{\tau} h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) + \log C_h(\mathbf{z}) \right] \tag{15}$$

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log C_h(\mathbf{z}) \right]. \tag{16}$$

Using the definition of C_h in Eq. (12) yields

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \int_{\mathcal{Z}} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}} \right]. \tag{17}$$

We assume a uniform marginal distribution $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We expand by $|\mathcal{Z}||\mathcal{Z}|^{-1}$ and estimate the integral by sampling from $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, yielding

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log |\mathcal{Z}| \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] \tag{18}$$

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \right] \right] + \log |\mathcal{Z}| \tag{19}$$

By inserting the definition $h = f \circ g$,

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[(f \circ g)(\tilde{\mathbf{z}})^\top (f \circ g)(\mathbf{z}) \right] \tag{20}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})\tau} \right] \right] + \log |\mathcal{Z}|, \tag{21}$$

we can identify the losses introduced in Proposition A,

$$= \mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uniform}}(f;\tau) + \log |\mathcal{Z}|, \tag{22}$$

which recovers the original alignment term and the uniformity term for maximimizing entropy by means of a von Mises-Fisher KDE up to the constant $\log |\mathcal{Z}|$. According to proposition A this equals

$$= \lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log |Z|, \tag{23}$$

which concludes the proof. □

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} \exp(\kappa \tilde{\mathbf{z}}^\top \mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau/\kappa}$.[3] Consider the model conditional distribution*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \quad \textit{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \, \mathrm{d}\tilde{\mathbf{z}}, \tag{24}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.

[3] Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h—in particular, for $h(\mathbf{z}) = \sqrt{\tau/\kappa}\mathbf{z}$. The global minimum of the cross-entropy between two distributions is reached if they match by value and have the same support. Thus, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}^*}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{25}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using that h maps from a unit hypersphere onto one with radius $\sqrt{\tau/\kappa}$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_{\mathrm{h}^*}(\mathbf{z}|\mathbf{z}) \tag{26}$$

$$C_p \exp(\kappa \mathbf{z}^\top \mathbf{z}) = C_h(\mathbf{z}) \exp(h^*(\mathbf{z})^\top h^*(\mathbf{z})/\tau) \tag{27}$$

$$C_p \exp(\kappa) = C_h(\mathbf{z}) \exp(\kappa) \tag{28}$$

$$C_p = C_h. \tag{29}$$

As the normalization constants are identical we get for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$\exp(\kappa \mathbf{z}^\top \tilde{\mathbf{z}}) = \exp(h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}})) \Leftrightarrow \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}}). \tag{30}$$

□

Lemma 1 (Kernel Density Estimators in the limit of unlimited samples)**.** *For $0 < \alpha, \kappa < \infty$, we have*

$$p(\mathbf{z}) = \frac{1}{\alpha} \int p(\tilde{\mathbf{z}}) \exp(-\alpha \|\mathbf{z} - \tilde{\mathbf{z}}\|) \mathrm{d}\tilde{\mathbf{z}} \qquad \mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^N \tag{31}$$

$$p(\mathbf{z}) = C(\kappa) \int p(\tilde{\mathbf{z}}) \exp(\kappa \mathbf{z}^\top \tilde{\mathbf{z}}) \mathrm{d}\tilde{\mathbf{z}} \qquad \mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{S}^{N-1}. \tag{32}$$

Proof. First, we consider $\mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^N$. We can define the Kernel Density Estimator (KDE) for which $\lim_{n\to\infty} \hat{p}_n(\mathbf{z}) = p(\mathbf{z})$ [33]. We use this result to obtain

$$\lim_{n\to\infty} \hat{p}_n(\mathbf{z}) = \lim_{n\to\infty} \frac{1}{\alpha} \sum_{i=1}^{n} \frac{1}{n} \exp(-\alpha \|\mathbf{z} - \mathbf{z}_i\|) \mathrm{d}\tilde{\mathbf{z}}, \quad \mathbf{z}_i \sim^{\text{iid}} p(\mathbf{z}) \tag{33}$$

$$= \frac{1}{\alpha} \int_{\mathbb{R}^N} p(\mathbf{z}) \exp(-\alpha \|\mathbf{z} - \tilde{\mathbf{z}}\|) \, \mathrm{d}\tilde{\mathbf{z}} = p(\mathbf{z}). \tag{34}$$

The result for $\mathbb{S}^{N-1}$ can be derived analogously. □

Lemma 2 (Matching conditionals imply bounded pushforward densities)**.** *Let p and q_{h} be conditional distributions as defined above. If they match, i.e. $p = q_{\mathrm{h}}$, and the marginal distribution is bounded, i.e. $p(\mathbf{z}) \leq p_{max} < \infty$, the pushforward density $p_{\#h}$ of $h(\mathbf{z}), \mathbf{z} \sim p(\mathbf{z})$ is bounded.*

Proof. We will show, that the KDE of the pushforward distribution $p_{\#h}$ is related to the normalization constant $C_h(\mathbf{z})$. Let $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \|\mathbf{z} - \tilde{\mathbf{z}}\|$ or $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = -\mathbf{z}^\top \tilde{\mathbf{z}}$ be the function used in the model conditional q_h and let C_δ be the appropriate normalization constant as defined in Lemma 1. For $p(\mathbf{z}) < \infty$, we get (cf. Lemma 1)

$$p_{\#h}(\mathbf{z}) = C_\delta \int p(\tilde{\mathbf{z}}) \exp(-\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))) \, d\tilde{\mathbf{z}} \tag{35}$$

$$\leq C_\delta \int p_{\max} \exp(-\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))) \, d\tilde{\mathbf{z}} \tag{36}$$

$$= C_\delta \, p_{\max} \, C_h(\mathbf{z}). \tag{37}$$

For matching distributions $q_h = p$, we have $C_h(\mathbf{z}) = C_p$ acccording to Proposition 3 and get

$$p_{\#h}(\mathbf{z}) \leq C_\delta \, p_{\max} \, C_p \tag{38}$$

Since both $p(\mathbf{z}) < \infty$ and $C_p < \infty$ by assumption, it follows that $p_{\#h}(\mathbf{z}) < \infty$, concluding the proof. □

Lemma 3. *Let h be as described in Proposition 1, such that it minimizes the cross-entropy. Then the empirical marginal distribution of the model (pushforward of p through h, denoted as $p_{\#h}$) is a uniform distribution, i.e. $p_{\#h}(\mathbf{z}) = const.$*

Proof. As per Proposition 1, the conditional distributions' normalization constants have the same value, i.e. $C_p = C_q$. However, as noted above (see Sec. A.2), the definition of C_q coincides with a von Mises-Fisher KDE of the empirical marginal distribution $p(h(\mathbf{z}))$ in the limit of infinite samples (Lemma. 1, 34). As this C_q equals $C_p = |\mathcal{Z}|^{-1}$, the empirical distribution $p(h(\mathbf{z}))$ is also a uniform distribution. □

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$. If $h : \mathcal{Z} \to \mathcal{Z}$ maintains the dot product up to a constant factor, i.e. $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an affine transformation.*

Proof. As h maintains the dot product up to a factor, this also holds true if one rotates the coordinate system by an arbitrary rotation matrix $\mathbf{R} \in \mathrm{SO}(N)$. Thus, we get

$$\forall \mathbf{R} \in \mathrm{SO}(N), \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \mathbf{R}^\top \mathbf{R} \tilde{\mathbf{z}} = h(\mathbf{R}\mathbf{z})^\top h(\mathbf{R}\tilde{\mathbf{z}}). \tag{39}$$

We consider the partial derivatives w.r.t. $\mathbf{z}$ and obtain:

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \tilde{\mathbf{z}} = \mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z}) h(\mathbf{R}\tilde{\mathbf{z}}). \tag{40}$$

We can recover the initial dot product by multiplying both sides of the equation with $\mathbf{z}^\top$ to obtain

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = \mathbf{z}^\top \mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z}) h(\mathbf{R}\tilde{\mathbf{z}}) \tag{41}$$

$$= h(\mathbf{R}\tilde{\mathbf{z}})^\top \mathbf{J}_h(\mathbf{R}\mathbf{z}) \mathbf{R}^\top \mathbf{z}. \tag{42}$$

From here, we take the partial derivative on both sides, this time w.r.t. $\tilde{\mathbf{z}}$, yielding

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z} = [\mathbf{R}\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}}) \mathbf{J}_h^\top(\mathbf{R}\mathbf{z}) \mathbf{R}^\top] \mathbf{z}. \tag{43}$$

Multiplying with $\mathbf{R}^\top$ from the left and defining $\mathbf{z}' := \mathbf{R}^\top \mathbf{z}$ gives

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}}) \mathbf{J}_h^\top(\mathbf{R}^2 \mathbf{z}')] \mathbf{z}'. \tag{44}$$

We define a transform from $(\mathbf{R}, \mathbf{z}, \tilde{\mathbf{z}})$ to $(\mathbf{a}, \mathbf{b}, \mathbf{z}')$: First, we select $\mathbf{R}$ and $\mathbf{z}$ s.t. $\mathbf{z}' = \mathbf{R}^T \mathbf{z}$ and $\mathbf{b} = \mathbf{R}\mathbf{z} = \mathbf{R}^2 \mathbf{z}'$. Then, we select $\tilde{\mathbf{z}}$ s.t. $\mathbf{a} = \mathbf{R}\tilde{\mathbf{z}}$. With this transform, we rewrite the aforementioned equation and obtain:

$$\forall \mathbf{a}, \mathbf{b}, \mathbf{z}' \in \mathcal{Z} : \kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{a}) \mathbf{J}_h(\mathbf{b})^\top] \mathbf{z}', \tag{45}$$

which can only be satisfied iff

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a}) \mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{I}. \tag{46}$$

By evaluating this expression for $\mathbf{a} = \mathbf{b}$ we get the

$$\forall \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{J}_h^{-1}(\mathbf{b}) \tag{47}$$

Inserting this property again in the previous expression yields

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a}) \kappa \mathbf{J}_h(\mathbf{b})^{-1} = \kappa \mathbf{I}, \tag{48}$$

and finally:

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z} : \mathbf{J}_h(\mathbf{a}) = \mathbf{J}_h(\mathbf{b}). \tag{49}$$

□

Taking all of this together, we can now prove Theorem 2:

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$ and the ground-truth marginal be uniform and the conditional a vMF distribution (cf. Eq. 2). If the mixing function g is differentiable and invertible, and if f is differentiable and minimizes the CL loss (1), then for fixed $\tau > 0$ and $M \to \infty$ we find that $h = f \circ g$ is affine, i.e. we recover the latent sources up to affine transformations.*

Proof. As f minimzes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ we can apply Theorem 1 to see that f also minimizes the cross-entropy between $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$. This means, we can apply Proposition 1 to show that the concatenation $h = f \circ g$ is an isometry with respect to the dot product. Finally, according to Proposition 2, h must then be a linear transformation on the hypersphere, i.e., a combination of permutations, sign flips and rotations. Thus, f recorvers the latent sources up to affine transformations, concluding the proof. □

A.4 Extension to (subspaces of) $\mathbb{R}^N$

Here, we show how one can generalize the theory above from $\mathcal{Z} = \mathbb{S}^{N-1}$ to $\mathcal{Z} \subseteq \mathbb{R}^N$. Under mild assumptions regarding the ground-truth conditional distribution p and the model distribution q_{h}, we prove that all minimizers of the cross-entropy between p and q_{h} are linear functions, if $\mathcal{Z} = \mathbb{R}^N$ or $\mathcal{Z}$ is a convex body. Note that the hypercube $[a_1, b_1] \times \ldots \times [a_N, b_N]$ is an example of such a convex body.

A.4.1 Assumptions

First, we restate the core assumptions for this proof. The main difference with the assumptions for the hyperspherical case above is that we assume different conditional distributions: instead of rotation-invariant von Mises-Fisher distributions, we use translation-invariant distributions (up to restrictions determined by the finite size of the space) of the exponential family.

Generative Process Let $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} \subseteq \mathbb{R}^N$ and $\mathcal{X} \subseteq \mathbb{R}^N$, with $\mathcal{Z} = \mathbb{R}^N$ or $\mathcal{Z}$ being a convex body (e.g. a hypercube). Further, let the marginal distribution fulfill $\operatorname{supp} p(\mathbf{z}) = \mathcal{Z}$. We assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is an exponential distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) e^{-\lambda \delta(\tilde{\mathbf{z}}, \mathbf{z})} \quad \text{with} \quad C_p(\mathbf{z}) := \int e^{-\lambda \delta(\mathbf{z}, \tilde{\mathbf{z}})} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{50}$$

where δ is a semi-metric, and $\lambda > 0$ a parameter controlling the width of the distribution. We make no further assumptions on the marginal distribution $p(\mathbf{z})$ except that it must be compatible with the conditional distribution:

$$p(\mathbf{z}) = \int p(\mathbf{z}, \tilde{\mathbf{z}}) \, \mathrm{d}\tilde{\mathbf{z}} = \int p(\mathbf{z}|\tilde{\mathbf{z}}) p(\tilde{\mathbf{z}}) \, \mathrm{d}\tilde{\mathbf{z}}. \tag{51}$$

Finally, we assume that during training one has access to samples from both of these distributions.

Model Let $\mathcal{Z}'$ be a subset of $\mathbb{R}^N$ that is a convex body and let $f : \mathcal{X} \to \mathcal{Z}'$ be the model whose parameters are optimized. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))} \quad \text{with} \quad C_q(\mathbf{z}) := \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{52}$$

where $C_q(\mathbf{z})$ is the partition function and δ is defined above.

A.4.2 Minimizing the cross-entropy

In a first step, we derive a property similar to Theorem 1, which suggests a practical method to find minimizers of the cross-entropy between the ground-truth p and model conditional q_{h}. For this, we suggest a modification of the contrastive loss Eq. (1).

Proposition A.4.2. *Let δ be a semi-metric. Consider the ground-truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\lambda \delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and the model conditional distribution*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))), \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))) \mathrm{d}\tilde{\mathbf{z}}. \tag{53}$$

Then the cross-entropy between p and q_{h} is upper bounded by

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}), q_{\mathrm{h}}(\cdot|\mathbf{z})\right] \leq \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \tag{54}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right]\right)\right] + \textit{const.}, \tag{55}$$

which can be implemented by sampling data from the accessible distributions. This bound becomes tighter, the more uniform *the marginal distribution p becomes. Note, the similarity to the InfoNCE objective in Eq. (1), where the distance function δ corresponds to the dot-product.*

Proof.

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}), q_{\mathrm{h}}(\cdot|\mathbf{z})\right] = -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}))\right]\right] \tag{56}$$

By inserting the definition of q_{h} one gets

$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(C_h^{-1}(\mathbf{z})) - \delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right]\right] \tag{57}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(C_h(\mathbf{z})) + \delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right]\right] \tag{58}$$

As $C_h(\mathbf{z})$ does not depend on $\tilde{\mathbf{z}}$ it can be moved out of the inner expectation value, yielding

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \log(C_h(\mathbf{z}))\right], \tag{59}$$

which can be written as

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log(C_h(\mathbf{z}))\right], \tag{60}$$

Inserting the definition of C_h gives

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\int \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right)\right]. \tag{61}$$

By using that p is a regular probability density and writing $p_{\max} = \max_{\mathbf{z}} p(\mathbf{z})$, the second term can be upper bounded with

$$\leq \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\int \frac{p(\tilde{\mathbf{z}})}{p_{\max}} \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right)\right], \tag{62}$$

with equality if the marginal p is a uniform distribution over $\mathcal{Z}$. Finally, this can be simplified as

$$= \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}\left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[\exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\right]\right)\right] - \log p_{\max}. \tag{63}$$

□

A.4.3 Cross-entropy minimizers are isometries

Now we show a version of Proposition 1, that is generalized from hyperspherical spaces to (subsets of) $\mathbb{R}^N$. We note that while providing an exact link between cross-entropy minimization and CL with marginals on $\mathbb{R}^N$ is beyond the scope of this work, this result can still serve as a stepping stone for generalizing our identifiability result on the hypersphere in future work.

Proposition 3 (Minimizers of the cross-entropy are isometries)**.** *Let δ be a semi-metric. Consider the conditional distributions of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))), \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})))\mathrm{d}\tilde{\mathbf{z}}, \tag{64}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathcal{L}_{\mathsf{CE}} = \mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then h is an isometry, i.e. $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \lambda\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h^*(\mathbf{z}), h^*(\tilde{\mathbf{z}}))$. Note, that this does not depend on the choice of $\mathcal{Z}$ but just on the class of conditional distributions allowed.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h, e.g. the identity. The global minimum of cross-entropy between two distributions is reached if they match by value and have the same support. Hence, if p is a regular density, q_{h} will be a regular density, i.e. q_{h} is continuous and has only finite values $0 \leq q_{\mathrm{h}} < \infty$. As the two distributions match, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{h^*}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{65}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using the property $\delta(\mathbf{z}, \mathbf{z}) = 0$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_{\mathrm{h}*}(\mathbf{z}|\mathbf{z}) \tag{66}$$

$$\Leftrightarrow \quad C_p^{-1}(\mathbf{z}) \exp(-\lambda\delta(\mathbf{z}, \mathbf{z})) = C_h^{-1}(\mathbf{z}) \exp(-\delta(h^*(\mathbf{z}), h^*(\mathbf{z}))) \tag{67}$$

$$\Leftrightarrow \quad C_p(\mathbf{z}) = C_h(\mathbf{z}). \tag{68}$$

As the normalization constants are identical, we obtain for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$\exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z})) = \exp(-\delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z}))) \Leftrightarrow \lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z})). \tag{69}$$

By introducing a new semi-metric $\delta' := \lambda^{-1}\delta$, we can write this as $\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta'(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z}))$, which shows that h is an isometry. □

A.4.4 Bijectivity of h

We proceed by showing an important property that will be needed later in Sec. A.5: the bijectivity of h. For this, we show that if the conditional probability distribtion q_{h}, which is implicitly defined by h, is a regular density function (i.e. $0 \leq q_{\mathrm{h}} < \infty$), then h is bijective. First, we start with introducing some existing concepts:

Definition 1. Let $\mathcal{M}, \mathcal{N}$ be manifolds. A map $h : M \to N$ is *proper* if for every compact set $S \in \mathcal{N}$ its preimage $h^{-1}(S)$ is compact in $\mathcal{M}$.

Proposition 4. *Let $\mathcal{M}$ and $\mathcal{N}$ be simply connected, oriented, d-dimensional $\mathcal{C}^1$-submanifolds of $\mathbb{R}^D$ ($D \geq d$), without boundary. Let $h : M \mapsto N$ be a proper $\mathcal{C}^1$-map such that the Jacobian determinant $|\mathbf{J}_h|$ never vanishes. Then h is bijective.*

Proof. See Theorem 2.1 in [35]. □

Many important manifolds fit the conditions of this proposition, including $\mathbb{R}^N$. We now prove that under mild conditions, any h, such that q_{h} as defined above is a regular density function, is proper and has a non-vanishing Jacobian determinant.

Proposition 5. *Let $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map, such that q_{h} as defined abvove is a regular density function; this also means that $\sup q_{\mathrm{h}} < \infty$. Then h is proper and has a non-vanishing Jacobian determinant.*

Proof. Suppose that the Jacobian of h vanishes for some $\mathbf{z} \in \mathcal{M}$. Then the inverse of the determinant of the Jacobian goes to infinity at this point and so does the density of $h(\mathbf{z})$ according to the well-known transformation of probability densities. By assumption, q_{h} must be a regular density function and, therefore, cannot be a delta distribution.

The mapping h is proper if pre-images of compact spaces are compact. According to the Heine–Borel theorem, compact subsets of $\mathbb{R}^D$ are closed and bounded. Additionally, a continuous mapping between $\mathcal{M}$ and $\mathcal{N}$ is also closed, i.e. pre-images of closed subsets are also closed [36]. In addition, it is well-known that continuous functions on compact spaces are bounded[4], concluding the proof. □

[4] See e.g. https://proofwiki.org/wiki/Continuous_Function_on_Compact_Space_is_Bounded

Finally, this is sufficient to prove that the map h is bijective, if q_{h} is a regular density:

Theorem 3. *Let $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map, such that q_{h} as defined abvove is a regular density function, and with $\mathcal{M}$, $\mathcal{N}$ defined as above. Then h is bijective.*

Proof. According to Proposition 5, h is proper and has non-vanishing Jacobians. Then, according to Proposition 4, h is bijective. □

This general property also holds for the special case of maps between $\mathbb{R}^N$:

Corollary 1. *Let $h : \mathbb{R}^N \to \mathbb{R}^N$ be a differentiable map, such that q_{h} as defined abvove is a regular density function as defined as above. Then h is bijective.*

A.5 Cross-entropy minimization identifies the ground-truth factors

Before we continue, let us recall a Theorem by Mazur and Ulam [37]:

Theorem 4 (Mazur-Ulam)**.** *Every bijective isometry between real normed spaces is affine.*

Proof. See Mazur and Ulam [37] or Nica [38]. □

Furthermore, let us recall an extension of that Theorem for convex bodies by Mankiewicz [39]:

Theorem 5 (Mankiewicz)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces and let $\mathcal{V}$ be a convex body in $\mathcal{X}$ and $\mathcal{W}$ a convex body in $\mathcal{Y}$. Then every isometry of between $\mathcal{V}$ and $\mathcal{W}$ can be uniquely extended to an affine isometry between $\mathcal{X}$ and $\mathcal{Y}$.*

Proof. See Mankiewicz [39]. □

By combining the properties derived before we can show that h is an affine function:

Theorem 6. *Let $\mathcal{Z}, \mathcal{Z}'$ be a convex bodies in $\mathbb{R}^N$ or the entire $\mathbb{R}^N$. If h minimizes the cross-entropy between p and q_{h} as defined in (4) and if the mixing function $\mathbf{g}$ is differentiable and invertible, then we find that $h = f \circ g$ is affine, i.e. we recover the latent sources up to affine transformations.*

Proof. According to Proposition 3 h is an isometry and q_{h} is a regular probability density function. Then, according to Corollary 1 h is also bijective. Finally, Theorem 4 says that h is an affine transformation. □

Note, that this result can be seen as a first step towards a generalized version of Theorem 2, as it is valid for $\mathcal{Z} = \mathbb{R}^N$ and allows a larger variety of conditional distributions. A missing step is to derive a connection between minimizing $\mathcal{L}_{\mathsf{contr}}$ and minimizing the cross-entropy, just as Theorem 1 does for the hyperspherical case. This will be addressed in future work.

A.6 Extended experiments

A.6.1 Implementation

Inspired by the experimental evaluation by Wang and Isola [18], we use a convex combination of the $\mathcal{L}_{\mathsf{align}}$ and $\mathcal{L}_{\mathsf{uniform}}$ losses. Given the severe mismatch between DSprites and our theory's assumptions, we generalized the $\mathcal{L}_{\mathsf{align}}$ loss function for more flexibility, resulting in the loss function

$$\mathcal{L}_{\mathsf{AU}}(\alpha, p) := \alpha \widetilde{\mathcal{L}}_{\mathsf{uniform}}(p) + (1-\alpha)\widetilde{\mathcal{L}}_{\mathsf{align}}(p), \tag{70}$$

with the components

$$\widetilde{\mathcal{L}}_{\mathsf{align}}(p) := \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}},\mathbf{z})\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[\|(f\circ g)(\mathbf{z}) - (f\circ g)(\mathbf{z})\|_p^p\right] \tag{71}$$

$$\widetilde{\mathcal{L}}_{\mathsf{uniform}}(p) := \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\mathbf{z}^-\sim p(\mathbf{z}^-)} \left[e^{-\|(f\circ g)(\mathbf{z}) - (f\circ g)(\mathbf{z})\|_p^p}\right]\right]. \tag{72}$$

A.6.2 Disentanglement Datasets

Standard datasets for disentanglement, most of which have been compiled by Locatello et al. [40], are limited in that the data generating process is independent and identically distributed (*i.i.d.*). This is problematic for evaluating the identifiability capabilities of contrastive methods, which require *positive pairs* to enforce alignment in latent space. Recent work in disentanglement, which make assumptions about pairs of instances in order to perform nonlinear demixing in an unsupervised fashion [24, 30, 31], introduced datasets in order to overcome the cited issues. We leverage these datasets for evaluation on more complex inputs.

UNI [30] Pairs of images are combined such that only k factors change, where $k \in \mathcal{U}\{1, D-1\}$ and D denotes the number of ground-truth factors, where the next value for each of the k factors is sampled uniformly from the set of possible values.

LAP [31] For each ground-truth factor, the first value in the pair is chosen i.i.d. from the dataset and the second is chosen by weighting nearby factor values using Laplace distributed probabilities. If all factors remain constant (no transition), then the sample is rejected because the pair would not result in any temporal learning signal.

NAT [31] For a given image pair, the position and scale of the sprite objects are set using measured values from adjacent time points for natural objects in YouTube-VOS [41, 42]. The sprite shapes are simple, like dSprites [43], and fixed for a given pair. The sprite orientations are fixed for the pair and are sampled uniformly from the same distribution as was used for dSprites. A version discretized to the granularity of dSprites was contributed, we refer to the discrete version as **ND** and the continuous version as **NC**.

KITTI [31] The dataset is composed of pedestrian segmentation masks from an autonomous driving vision benchmark KITTI-MOTS [44], with natural shapes and continuous natural transitions.

A.6.3 Extended Disentanglement Results

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
Ada-GVAE (UNI)*	92.3	84.7	26.6	47.9	91.3	7.4
SlowVAE (UNI)*	90.4 (3.3)	81.35 (7.6)	35.7 (8.3)	52.1 (6.5)	87.6 (2.0)	5.1 (1.4)
$\mathcal{L}_{\mathsf{AU}}(0.9999, 1)$ (UNI)	82.5 (3.4)	62.9 (3.8)	7.8 (1.9)	22.2 (2.9)	98.8 (0.7)	4.9 (1.5)
$\mathcal{L}_{\mathsf{AU}}(0.9, 1)$ (UNI)	80.2 (0.5)	61.0 (0.3)	20.5 (0.2)	42.4 (0.4)	100.0 (0.0)	7.6 (0.1)
SlowVAE (LAP)*	100.0 (0.0)	98.32 (2.5)	27.8 (7.9)	65.3 (3.1)	97.0 (1.5)	6.1 (2.6)
$\mathcal{L}_{\mathsf{AU}}(0.999, 1)$ (LAP)	100.0 (0.0)	96.0 (3.6)	24.9 (2.5)	45.6 (0.8)	92.9 (0.1)	9.6 (2.5)
$\mathcal{L}_{\mathsf{AU}}(0.99, 1)$ (LAP)	100.0 (0.0)	94.5 (3.3)	17.4 (3.7)	54.0 (1.8)	94.7 (1.4)	7.0 (0.6)
$\mathcal{L}_{\mathsf{AU}}(0.9, 1)$ (LAP)	100.0 (0.0)	89.5 (1.7)	18.1 (4.6)	54.4 (0.5)	90.5 (0.6)	6.7 (0.7)

Table 2: **DSprites.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (rows highlighted with * are from [31]).

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP	MCC
β-VAE	78.1 (3.0)	60.6 (6.0)	4.6 (1.9)	10.3 (1.8)	87.8 (2.3)	2.1 (1.0)	41.7 (3.4)
SlowVAE	82.6 (2.2)	76.2 (4.8)	11.7 (5.0)	18.9 (5.5)	88.1 (3.6)	4.4 (2.3)	52.6 (4.1)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$	72.2 (6.3)	62.1 (4.8)	9.4 (2.2)	18.1 (2.5)	96.5 (0.6)	5.1 (0.5)	33.4 (5.1)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$	55.7 (5.1)	44.1 (4.3)	1.7 (1.0)	3.2 (0.5)	91.2 (3.0)	0.7 (0.4)	22.4 (4.1)

Table 3: **Discrete Natural Sprites.** Mean (s.d.) performance levels over 10 random seeds.

Model (Data)	MCC
β-VAE	42.6 (4.7)
SlowVAE (C)	49.1 (4.0)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$	41.7 (5.3)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$	25.8 (4.6)

Table 4: **Continuous Natural Sprites**. Mean (s.d.) over 10 random seeds.

Model (frame separation)	MCC
β-VAE	62.7 (7.1)
SlowVAE (Δ t=1)	66.1 (4.5)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$ (Δ t=1)	77.1 (1.7)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$ (Δ t=1)	64.8 (2.4)
SlowVAE (Δ t=5)	79.6 (5.8)
$\mathcal{L}_{\mathsf{AU}}(0.5, 1)$ (Δ t=5)	79.6 (2.6)
$\mathcal{L}_{\mathsf{AU}}(0.5, 2)$ (Δ t=5)	68.3 (2.6)

Table 5: **KITTI Masks**. Mean (s.d.) over 10 random seeds.